CS224C: NLP for CSS

Computational Basics

Diyi Yang

Stanford CS

Announcements

Please sign up for Presentation/Scribe by this Friday (5pm PT)

Deadline updates for two classes:

Reading response for Jan 17th, due on Tuesday, Jan 17th 8am PT

Reading response for Feb 21th, due on Sunday, Feb 19th 5pm PT

See Canvas email for computing credits

Example reading response and presentation can be found on Canvas.

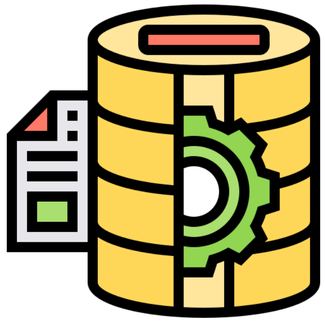
Clarification regarding Presentation

10~15 pages of slides

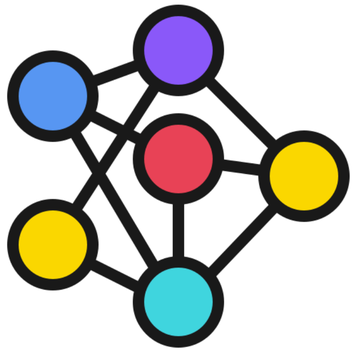
Only 10% of your grade

Attributes	CS224C NLP for Computational Social Science	CS224U: Natural Language Understanding	CS224N: Deep Learning for NLP	CS124: From Language to Information	CS384: Ethical & Social Issues in NLP
Audiences	Undergrad, Grad, and Non-CS major	Undergrad, Grad	Undergrad, Grad	Undergrad	Grad
Suggested Prerequistes	CS106B or equivalent	One of LINGUIST 180/280, CS 124, CS 224N, or CS 224S	Calculus and linear algebra; CS124, CS221, or CS229.	CS106B	CS224U or CS224N
Evaluation	Project, Quiz, Reading	Homework, Quiz, Project	Homework, Project	Programming homework, quiz, midterm	Homework, Project
Keywords	Applications in NLP Social science	Hands-on NLP Linguistics	Advanced NLP Deep Learning	Introduction to NLP, IR Social Networks	Ethics in NLP
Format	Pre-recorded Lectures Discussion	Lectures, Working Sessions, Podcast	Lectures Working Sessions	Flipped Class	Discussion
Interdisciplinary	*****	***	*	***	*****
Example Topics	NLP basics, cause inference, hypothesis testing, social influence, prosocial behavior, stigma/social movement	Word embedding, BERT, rational speech acts model, analysis methods in NLP, neural IR	Word vectors, language model, neural networks, parsing, pretraining, prompting, QA	Logistic regression, sentiment, IR, neural networks, chatbots, recommender systems, Pagerank and networks	Bias in NLP data and models, privacy, toxicity/abuse, fairness, stereotypes, propaganda

Computational Social Science in a nutshell



Data

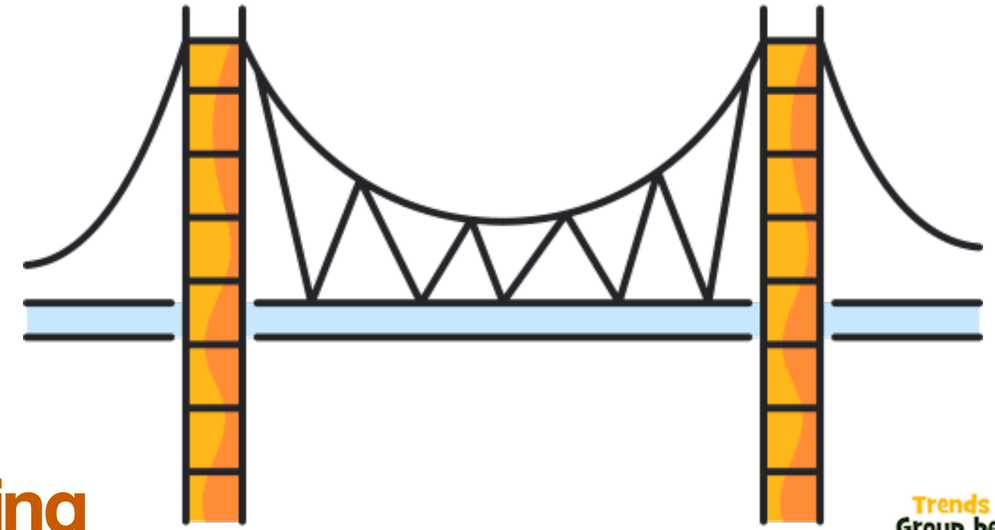


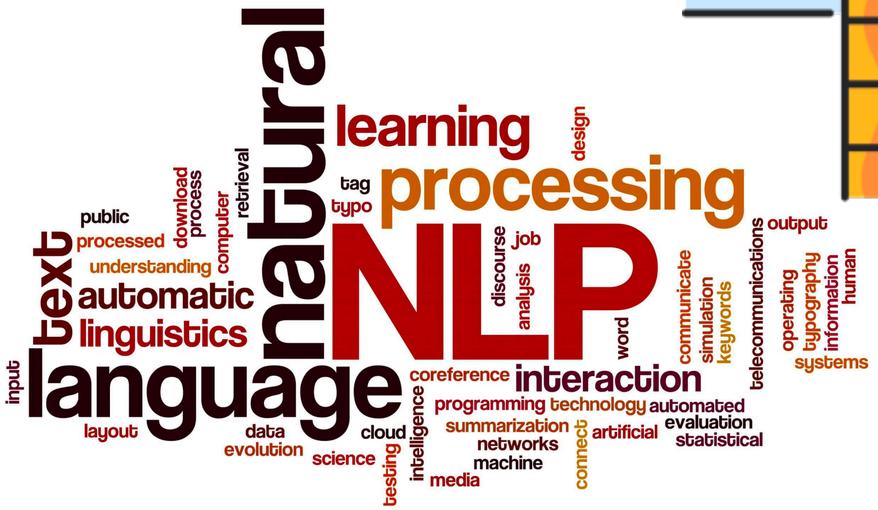
Algorithm

Problem

Knowledge

Impact



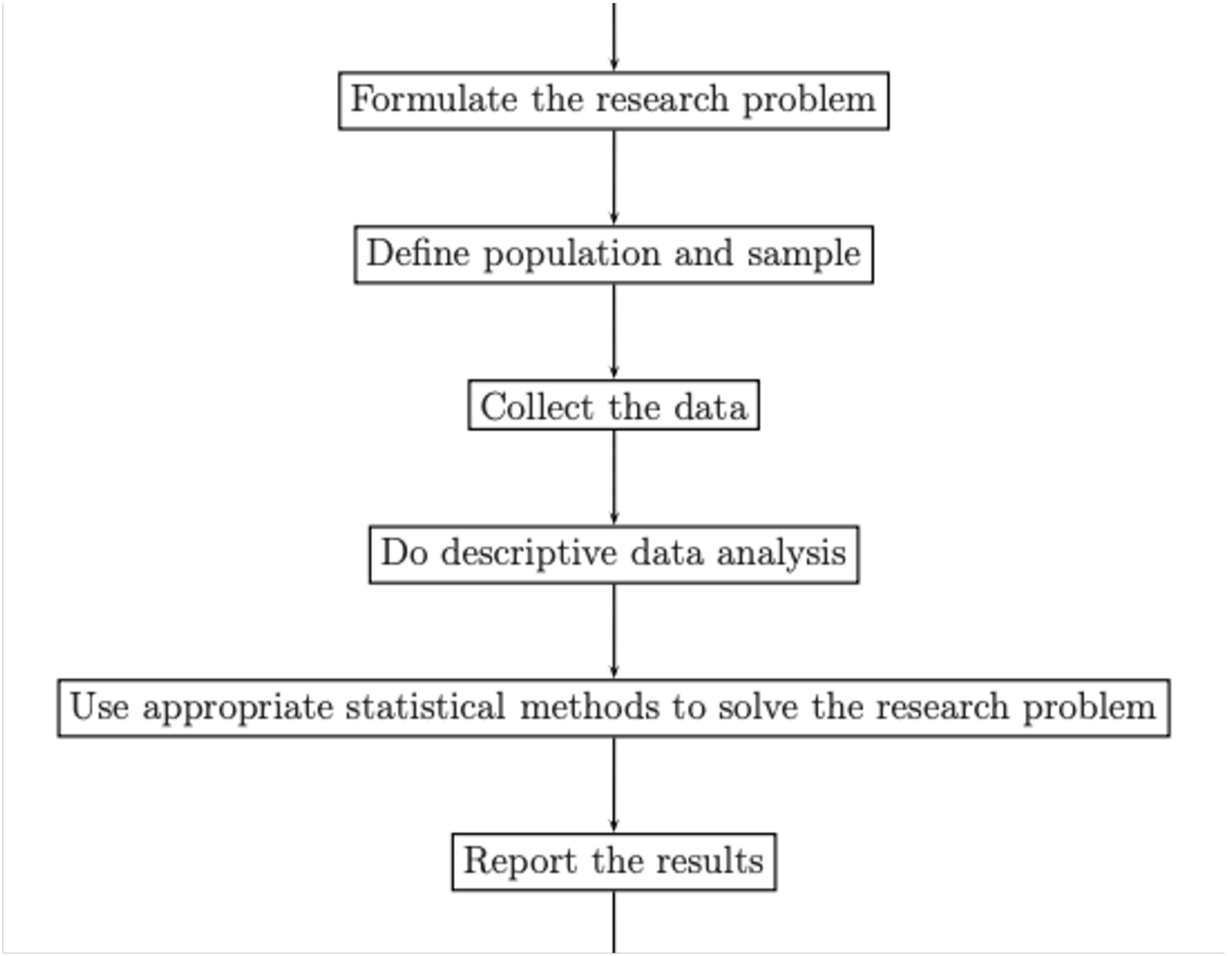

natural
learning
processing
NLP
language
text
automatic
linguistics
interaction
input
public
processed
understanding
download
process
computer
retrieval
tag
typo
design
discourse
analysis
job
word
communicate
simulation
keywords
telecommunications
output
operating
typography
information
human
systems
coreference
programming
technology
automated
evaluation
statistical
summarization
networks
machine
media
artificial
connect
intelligence
testing
cloud
data
evolution
science
layout


SOCIAL SCIENCES
AND HUMANITIES
Sociology
Anthropology
Infant
Injustice
Parent
Research
Design
Fashion
Food preparation
Food security
Regulations
Communicate
Inquiry
Garments
Behaviour
Cognitive
Mental Health
Nutrition
Trends
Group behaviour
Investigate
Psychology
Bias
Assess
Analyze
Children
Thesis
Metabolism
Fabrics
Human origins
Human mind
Equality
Family
Record
Recipe
Nutrients
Identity
Secondary source
Culture
Primary source

Lecture Overview

- Classification
- Regression
- Clustering
- Big Data + CSS

Computational Framework



Use of Classification or Regression

Two major uses of supervised classification/regression

Prediction:

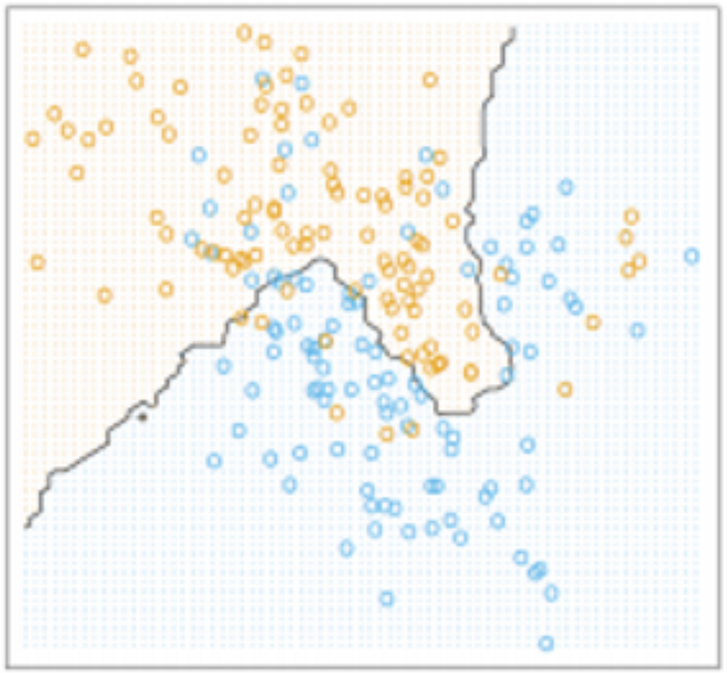
Train a model on a sample of data (x, y) to predict for some new data x'

Interpretation or Explanation:

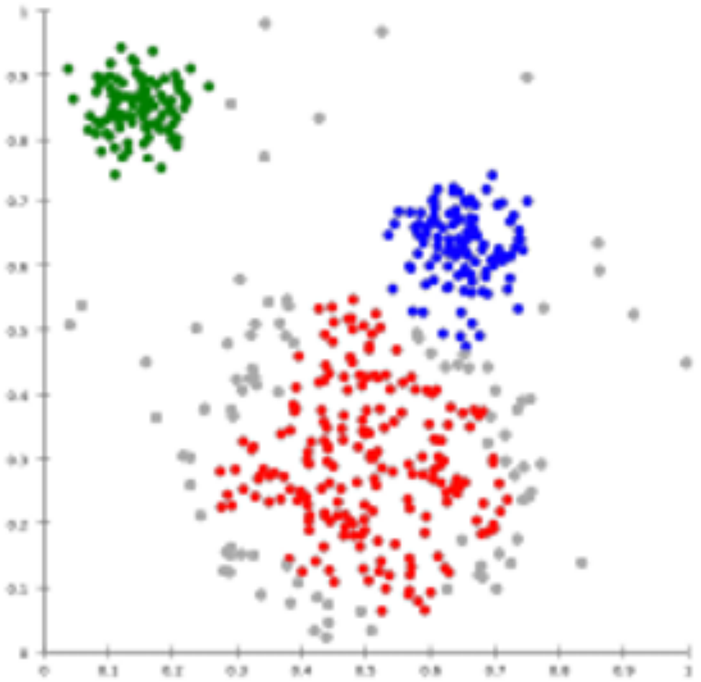
Train a model on a sample of data (x, y) to understand the relationship between x and y

Slide content credit to David Bamman

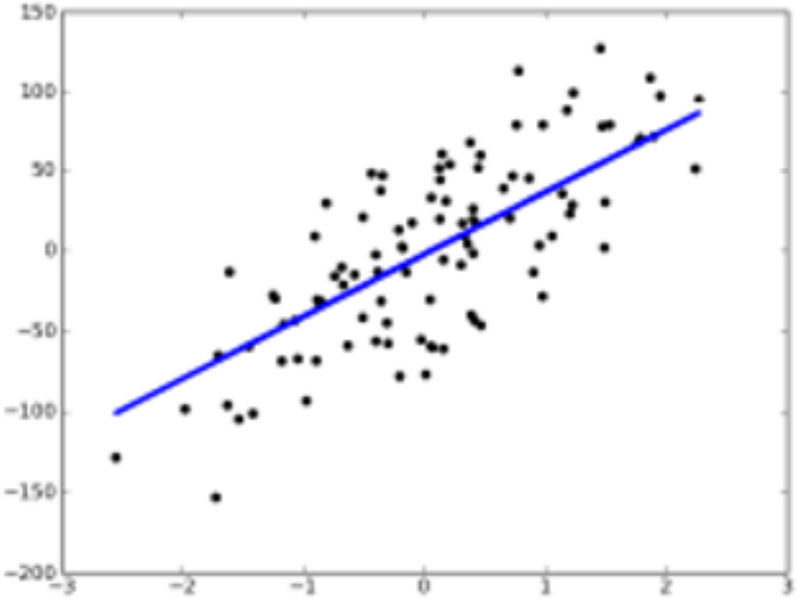
Common Methods



Classification



Clustering



Regression

Classification

A ***mapping*** h from input data x (drawn from instance space X) to a label y from some enumerable output space Y

X = set of all documents
Y = {English, Mandarin, Greek, …}
x = a single document
y = ancient Greek

Reviews and Ratings

Reviewed: October 24, 2022

8.0

Lovely little spot to spend some time. Very grateful for the clean, stylish room after travelling.

😀 · Beautiful, spacious room - after 22 hours of travel and a botched flight, I cried with happiness when I arrive.
Large, comfy bed - I didn't want to get out.
Friendly, helpful staff - especially the bar staff.
Accessible, enjoyable bar - open to late with a large selection of drinks.
Lots of room to sit by the pool. With a spa too. As well as on the foreshore in hammocks. Very laidback, enjoyable environment. I happily spent the day here, relaxing before a friends wedding. I loved walking along the foreshore footpath, following the shoreline and walking past the other hotels.

☹ · The beach wasn't an inviting swim, though a beautiful backdrop - which is not a fault of the hotel's. But flagging in case you're romanticising a beach swim; the hotel pool is better.
Watch: hidden costs. This might be normal/acceptable in non-Australian countries but I was caught off guard. There's the room cost, then there's the taxes (which booking.com tends to include in their final price), and THEN the hotel has a 'resort fee'. Which allows for 'amenities' access - which I find a bit "on the nose", the 'amenities' is what you automatically have access to when you book a room... but i guess some countres/states prefer a staggered bill...

IMDb Charts

IMDb Top 250 Movies

IMDb Top 250 as rated by regular IMDb voters.

Showing 250 Titles

Sort by: Ranking

Rank & Title	IMDb Rating	Your Rating
1. The Shawshank Redemption (1994)	★ 9.2	☆
2. The Godfather (1972)	★ 9.2	☆
3. The Dark Knight (2008)	★ 9.0	☆
4. The Godfather Part II (1974)	★ 9.0	☆
5. 12 Angry Men (1957)	★ 9.0	☆
6. Schindler's List (1993)	★ 8.9	☆

Some Text Classification Applications

Task	x	y
Language identification	text	{English, Mandarin, Greek, …}
Spam classification	email	{spam, not spam}
Authorship attribution	text	{jk rowling, james joyce, …}
Genre classification	novel	{detective, romance, gothic, …}
Sentiment classification	text	{positive, negative, neutral, mixed}

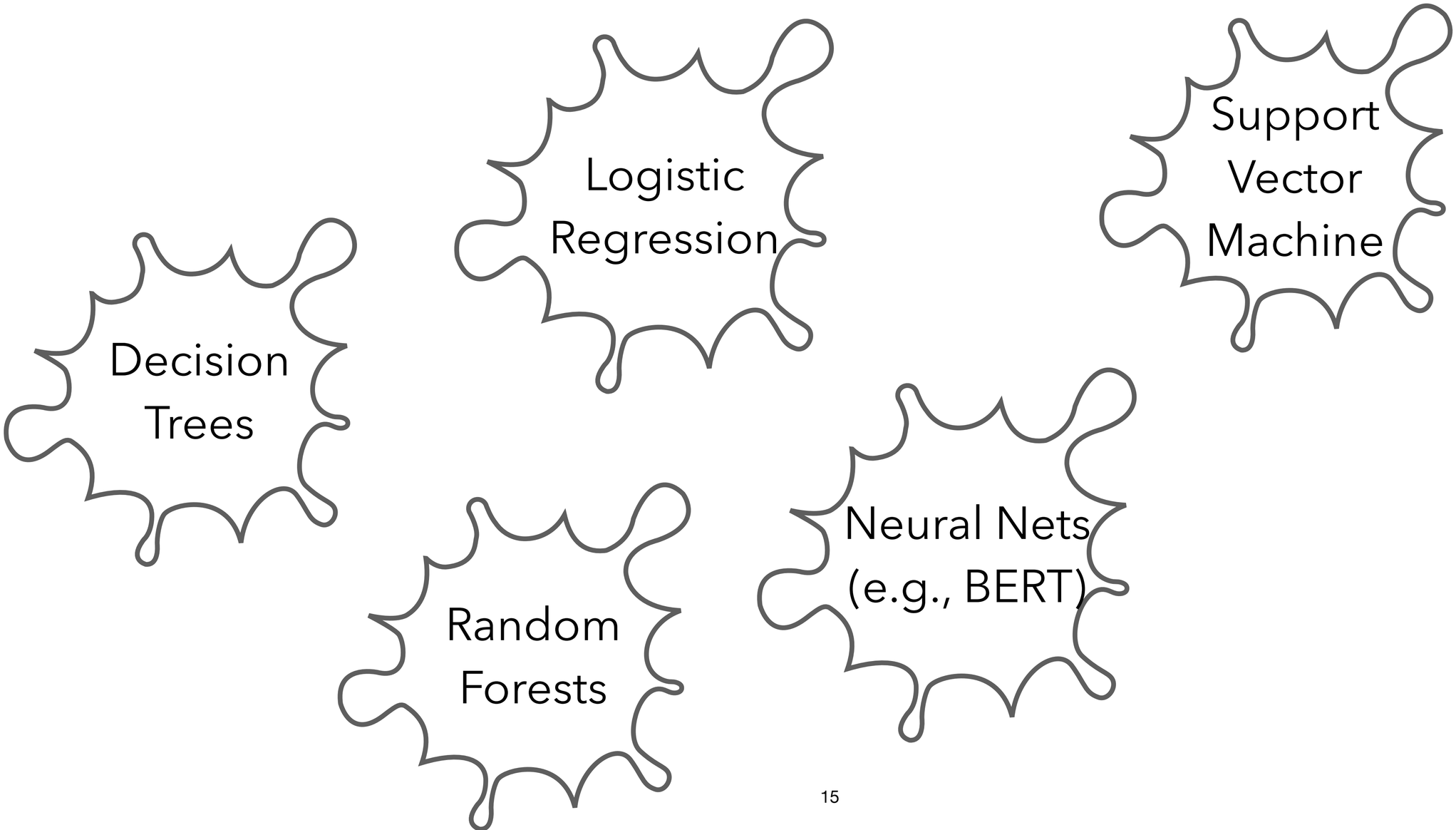
How to Perform Classification?

If we have a lot of data points (x, y)

Rule-based approaches

Supervised learning

Lots of Model Choices



Model Differences

Binary Classification

One out of 2 labels applied to a given x

Multiclass Classification

One out of N labels applied to a given x

Multilabel Classification

Multiple labels apply to a given x

Slide content credit to David Bamman

Recognizing A Classification Problem

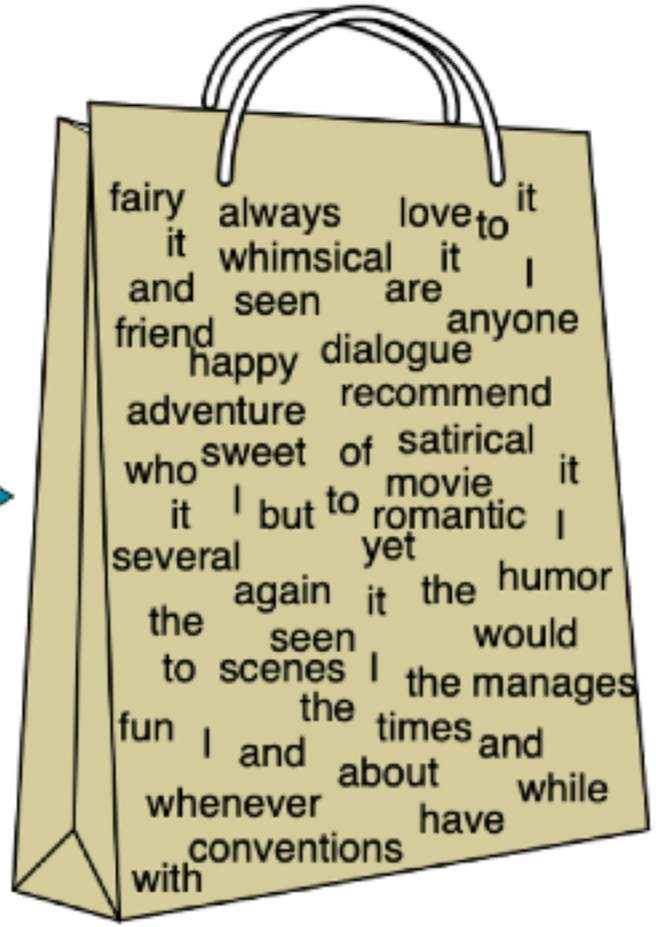
Can you formulate your question as a choice among some possible classes?

Can you create (or find) labeled data that marks that choice for a bunch of examples? Can you make that choice?

Can you create features that might help in distinguishing those classes?

Bag of Words

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Beyond the Bag of Words

Some linguistic phenomena require going beyond the bag-of-words:

- *That's not bad for the first day*
- *This is not the worst thing that can happen*
- *It would be nice if you acted like you understood*
- *This film should be brilliant. The actors are first grade. Stallone plays a happy, wonderful man. His sweet wife is beautiful and adores him. He has a fascinating gift for living life fully. It sounds like a great plot, however, the film is a failure.*

Applying Text Classification

The “raw” form of text is usually a sequence of characters

Converting this into a meaningful feature vector x requires a series of design decisions, such as tokenization, normalization, and filtering

Regression

Regression

A mapping from input data x (drawn from instance space X) to a point y in R

R: the set of real numbers

x = the empire state building

y = 17444.5625″

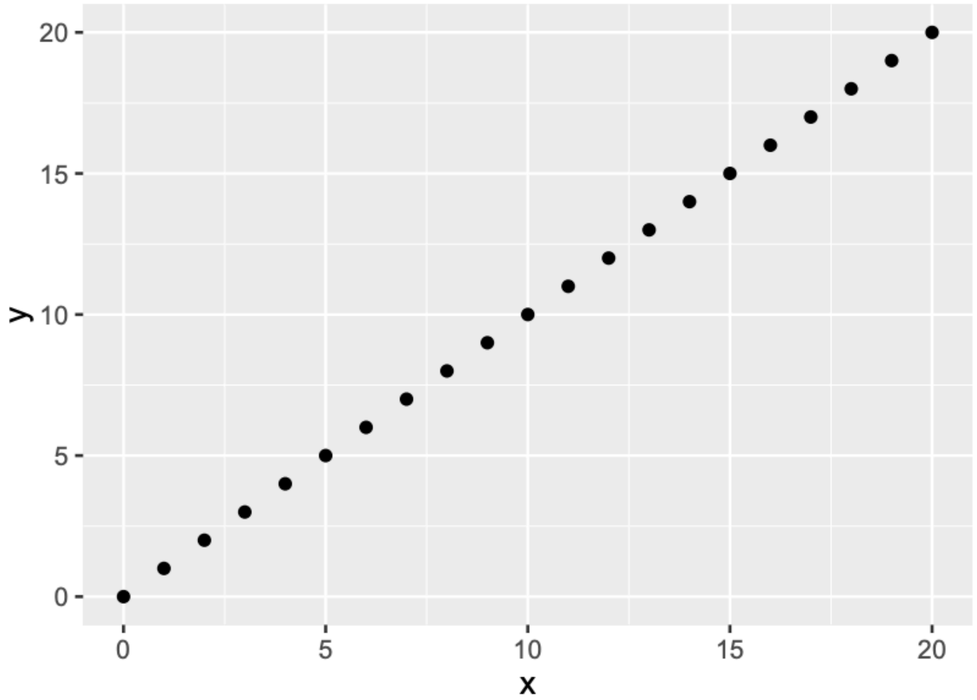
Slide content credit to David Bamman

Linear Regression

Suppose we have n data points. For each data point i, we observe

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$$

Linear regression states that $\hat{y}_i = \sum_{i=1}^{F} x_i \beta_i$



Slide content credit to David Bamman

Regression for Social Sciences

Regression for Social Sciences

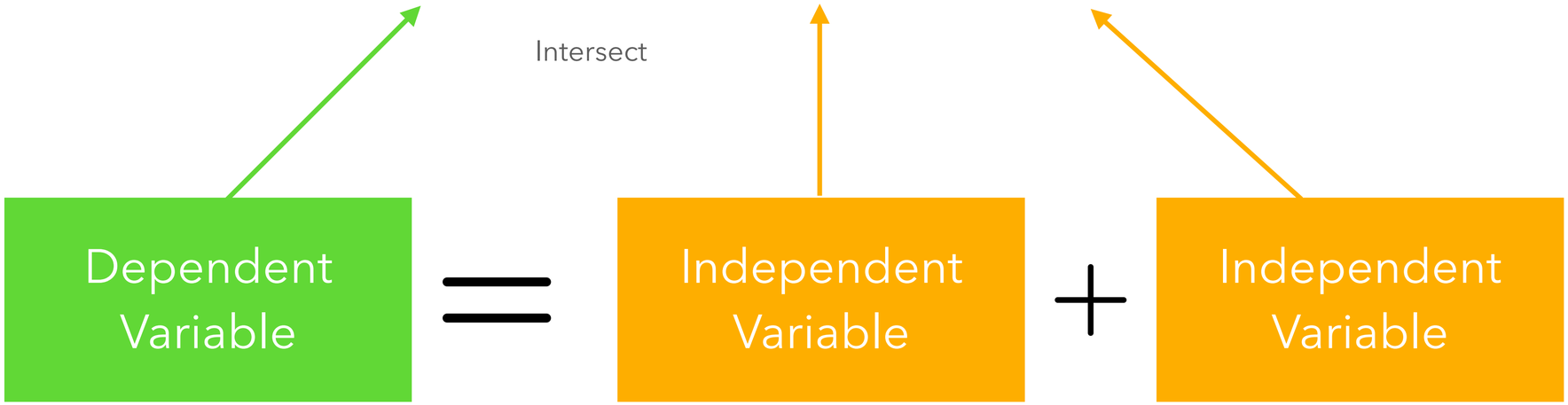
Regression analysis is a very useful tool for social sciences

- Understand the relationship between variables, adjusting for other potential confounders
- Predict the value of one variable based on others

In Other Terminology

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Intersect



How good is the Fit?

Mean squared error (MSE) $\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$

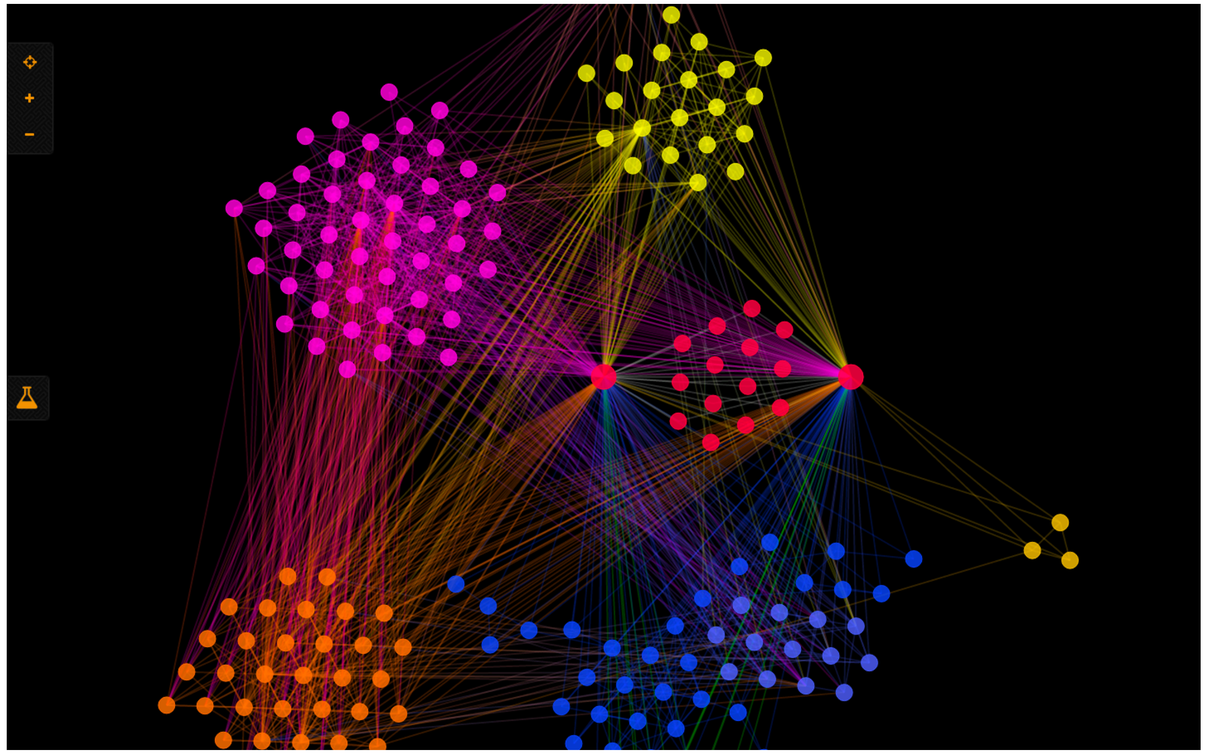
Mean absolute error (MAE) $\frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$

Clustering

Clustering

Group a set of data points into a number of clusters, so that

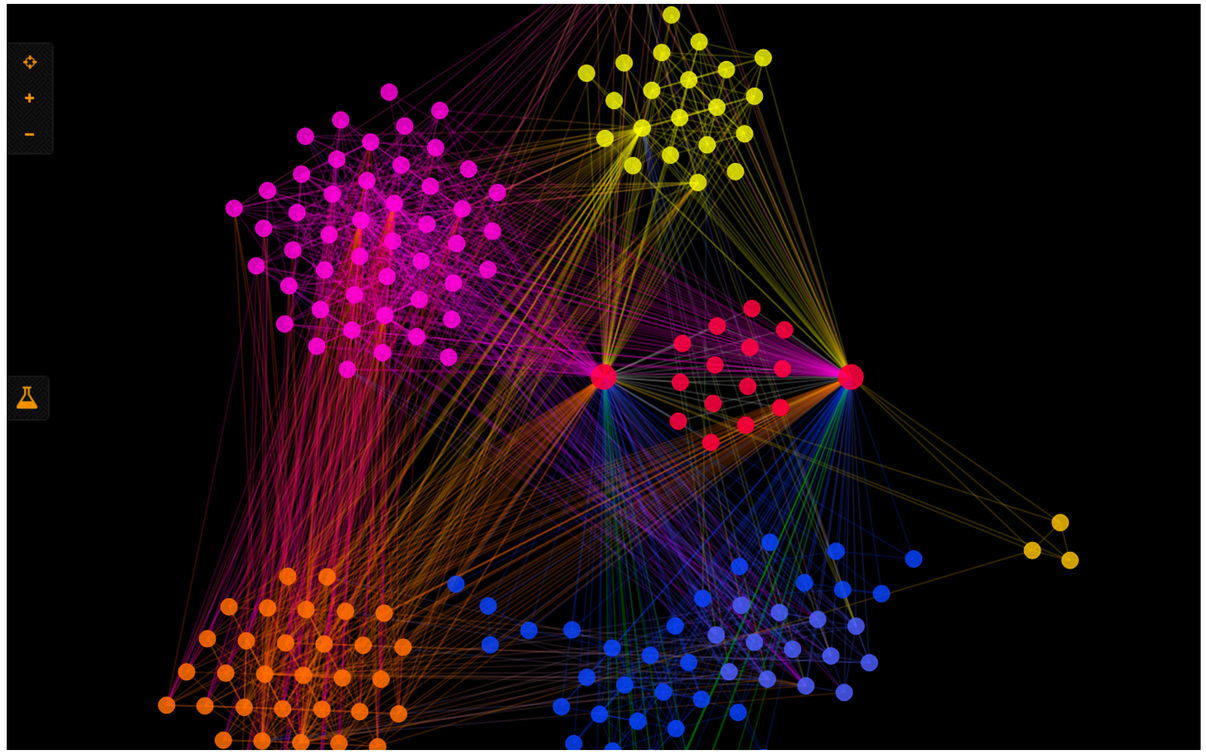
- Data points in the same cluster are similar to each other
- Data points in different clusters are dissimilar



https://graphalchemist.github.io/Alchemy/images/features/cluster_team.png

Clustering

Finding structures in data, using just X



https://graphalchemist.github.io/Alchemy/images/features/cluster_team.png

What are Structures?

Partitioning a group of data point into K disjoint sets (K-means clustering)

Assigning X to hierarchical structures (Hierarchical clustering)

Assigning X to partial membership in K different sets (Graphic models, GMM)

Learning a representation of x that puts similar data points closer to each other (Deep learning)

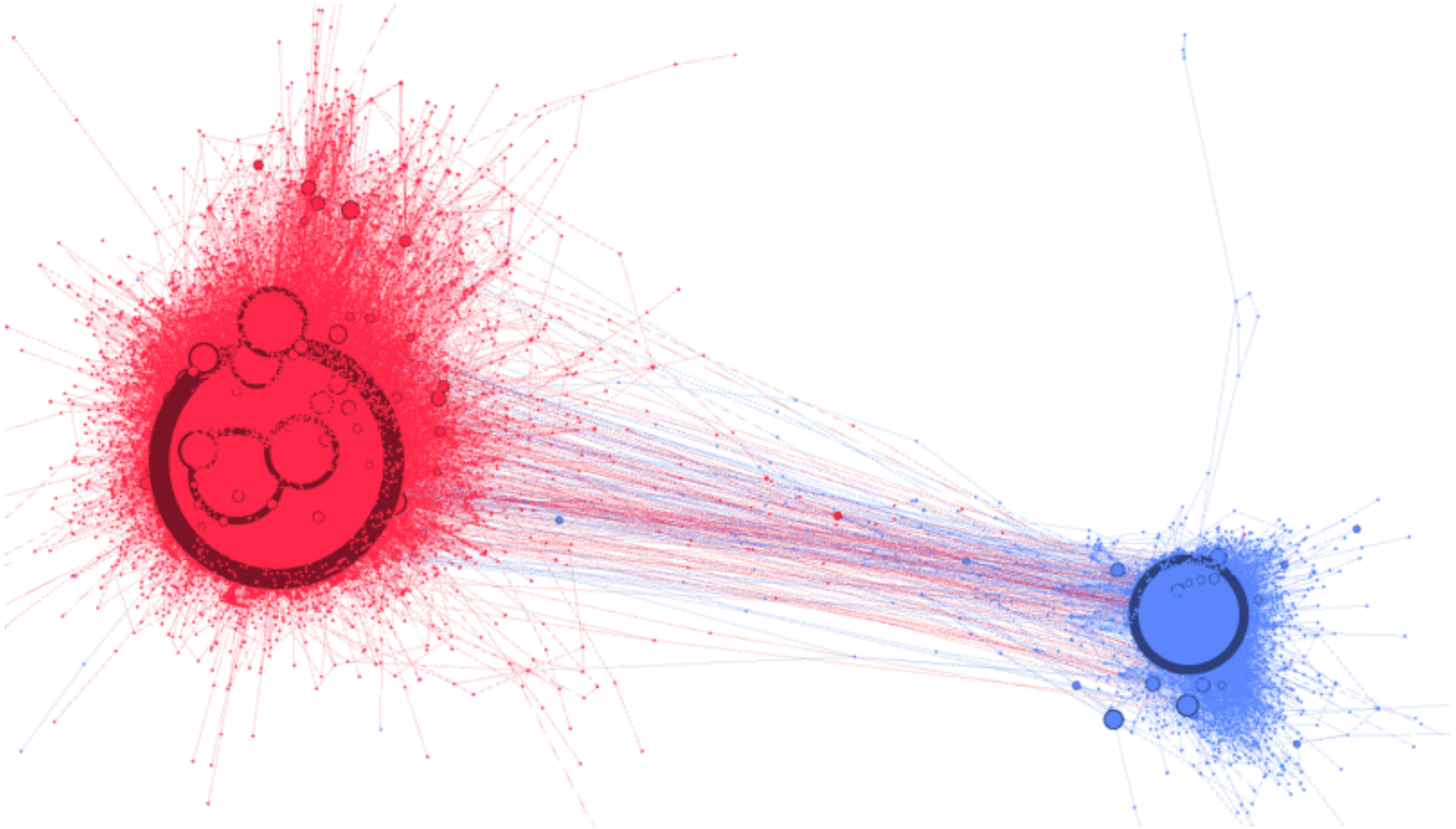
Slide content credit to David Bamman

Why and when do I need clustering?

Discovering interesting or unexpected structures can be useful for hypothesis generation

Unsupervised learning generates alternative representation **as features** for some subsequent supervised models

Slide content credit to David Bamman



The structure of the White Helmets discourse has two clear clusters of accounts—a pro-White Helmets cluster that supports the organization and an anti-White Helmets cluster that criticizes them, using Twitter conversations.

Wilson, Tom, and Kate Starbird. "Cross-platform disinformation campaigns: lessons learned and next steps." Harvard Kennedy School Misinformation Review 1, no. 1 (2020).

Key Design Choices for Clustering

How to **represent** each data point?

How to calculate the **similarity** between data points?

What is the **number of clusters** to use?

How can we **evaluate** the resulting clusters?

Is it a classification/regression/clustering problem?

I want to predict a star value {1,2,3,4,5} for a product review

I want to find all of the texts that have allusions to Paradise Lost

I want to predict the stock price

I want to tell which team will win

I want to associate photographs of cats with animals in a taxonomic hierarchy

I want to reconstruct an evolutionary tree for languages

Slide content credit to David Bamman

Computational Social Science in the Age of Big Data

danah boyd and Kate Crawford (2012), “Critical Questions for Big Data,” Information, Communication and Society

1 “Big data” changes the definition of knowledge

How do computational methods/quantitative analysis pragmatically affect epistemology?

Restricted to what data is available (twitter, data that’s digitized, google books, etc.). How do we counter this in experimental designs?

Establishes alternative norms for what “research” looks like

2 Claims to objectivity and accuracy are misleading

Data collection, selection process is subjective, reflecting belief in what matters.

Model design is likewise subjective

- model choice (classification vs. clustering etc.)
- representation of data
- feature selection

Claims need to match the sampling bias of the data

3 Bigger data is not always better data

Uncertainty about its source or selection mechanism [Twitter, Google books]

Appropriateness for question under examination

How did the data you have get there?

Are there other ways to solicit the data you need?

Remember **the value of small data**: individual examples and case studies

4 Taken out of context, big data loses its meaning

A representation (through features) is a necessary approximation; what are the consequences of that approximation?

Example: quantitative measures of “tie strength” and its interpretation

5 Just because it is accessible does not make it ethical

Anonymization practices for sensitive data (even if born public)

Accountability both to research practice and to subjects of analysis

6 Limited access to big data creates new digital divides

Inequalities in access to data and the production of knowledge

Privileging of skills required to produce knowledge

What's Next?

Mike Hardy and **Jiner Zheng** leading the discussion on Social Influence - Emotion Contagion next Tuesday!

Sign up for presentation/scribes!

Pre-recorded video on **Working with Text Data** will be released on Jan 17th.